Estimation of human endogenous retrovirus activities from expressed sequence databases

Merja Oja[1,2], Jaakko Peltonen[2,1], and Samuel Kaski[2]

[1] University of Helsinki, Department of Computer Science, FI-00014 University of Helsinki, Finland

[2] Helsinki University of Technology, Adaptive Informatics Research Centre, FI-02015 TKK, Finland

Abstract. Human endogenous retroviruses (HERVs) are remnants of ancient retrovirus infections and now reside within the human DNA. Recently HERV expression has been detected in both normal tissues and diseased patients. However, the activities (expression levels) of individual HERV sequences are mostly unknown. In this work we introduce a generative mixture model, based on Hidden Markov Models, for estimating the activities of the individual HERV sequences from databases of expressed sequences. We determine the relative activities of 91 HERVs; the majority of their activities were previously unknown. We also empirically justify a faster heuristic method for HERV activity estimation.

1 Introduction

Human endogenous retroviruses (HERVs) are remnants of ancient infections by retroviruses that have been fixed to human DNA. HERV sequences form 8% of the human genome. Retroviruses can move and copy their DNA to other locations in the genome; such copying eventually yields several mutated versions of the original virus. A group of such sequences is called a family, and may contain hundreds of very similar sequences.

HERVs are interesting for two reasons: they can express viral genes in human tissues, and their presence in the genome may affect the function of nearby human genes. Retroviral activity might cause disease; retroviral mRNAs have been detected in schizophrenia, autoimmune diseases and cancer [1]. In addition, a few retroviral genes have adopted functions beneficial to the human host [2].

In this paper we analyze the activity of individual HERV sequences. Being able to do so is vital for analyzing their control mechanisms and their possible roles in diseased and normal cell functions.[3] Most previous studies of HERV expression report activities only for HERV families (e.g. [3]); the only exceptions we know of are [4] where a small test for individual HERVs of one family was

[3] The activity of a HERV is influenced by control elements surrounding it in the DNA. To analyze this mechanism, we need to know which control elements are related to activity. The elements can be different around each individual HERV, so we need to know which individual HERVs are active.

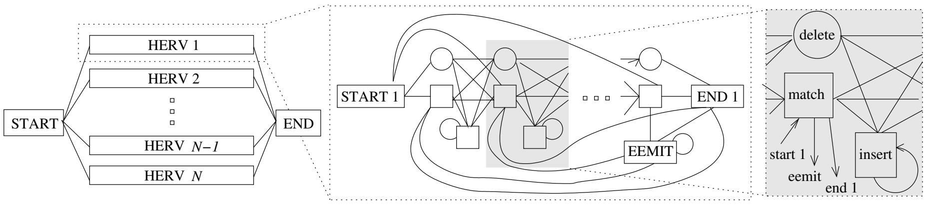

Fig. 1. The structure of the HMM mixture. The shaded box is the basic block of the sub-HMM and is repeated length-2 times. It is identical in all sub-HMMs; only the emission distribution of the match state varies between blocks, according to the HERV sequence each sub-HMM corresponds to. EEMIT-state emits the low-quality end part.

done with a heuristic method and [5] where HERVs are searched from gene mRNAs but activities are not compared across HERVs.

To find evidence of HERV expression, we use a large public database of expressed sequence tags (ESTs) which are short and noisy mRNA samples. The amount of ESTs available from a particular HERV is evidence of its activity (expression level). However, the noise level in ESTs can be larger than the sequence differences within a family, so it may be hard to determine exactly which HERV an EST stems from. We introduce a generative mixture model to model the uncertainty in the EST to HERV matching. The model learns the relative activities of the HERVs from EST sequence data.

2 Methods and Data

We introduce a generative mixture model for the set of EST sequences. The model is designed to mimic actual EST generation from HERVs; each mixture component is a Hidden Markov Model (HMM) for ESTs from a particular HERV. Such a HMM generates data that roughly matches a subsequence of the source HERV, but with mismatches, insertions, deletions, and a low-quality end part.

The mixture corresponds to one large HMM where the first transition chooses one of the N HERV-specific sub-HMMs (see Fig. 1). We use the Baum-Welch algorithm to learn the whole mixture. The learned probabilities of the first transition (the mixture weights) are estimates of the HERV activities.

We constrain the model complexity by sharing parameters. Each match state corresponds to a nucleotide in a HERV sequence: the probabilities of emitting the 'correct' HERV nucleotide or a mismatch are the same for all match states. Other emission and transition parameters are shared between all the basic blocks (shown in Fig. 1) of all the sub-HMMs.

In our pilot experiments we use 91 sequences randomly selected from three HERV families. The HERVs were detected from the human genome by the program RetroTector[4]. ESTs matching the HERVs were searched from the dbEST

[4] RetroTector is a program used for detecting retroviral sequences in genomes. It was developed by Jonas Blomberg and Göran Sperber at Uppsala University [6].

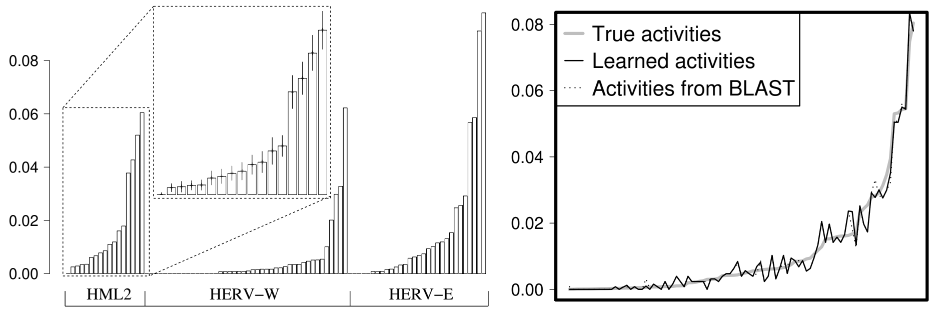

Fig. 2. Left: Real data. Relative activities of the 91 HERVs; crosses in the subfigure are means and standard deviations from resamples (see text). Right: Generated data. Activities learned by the HMM and the BLAST approach are compared to the ground truth.

database [7] with BLAST [8].[5] We used a strict match threshold of E-value 10^{-60} in BLAST. This yielded 1173 ESTs that match at least one HERV, with 9332 total matches.

The model is tested also on generated data. We generated ESTs from the 91 HERVs using our HMM model.[6] Then we used BLAST to match the HERVs against these ESTs and followed the same procedure as with real data. This yielded 1269 ESTs that match at least one HERV, with 8433 total matches.

We make the HMM training time reasonable by two heuristics: Only HERV-EST pairs returned by BLAST are used (others get zero probability), and we compare the EST and HERV sequences only close to the BLAST match area.[7]

3 Results

The method is able to estimate relative activities of the HERVs. The activity profile for the real data is in Fig. 2 (left). Each family has only a few highly active HERVs and many that are only slightly active. The activity of most of them was earlier unknown.

We estimated the reliability of the results with a bootstrap-like method. The EST data was resampled with replacement 10000 times, and activities were reoptimized for each replicate; other parameters were kept fixed. Fig. 2 (left) shows the mean and standard deviations of these replicates for each HERV in one

[5] The ESTs in dbEST are measured from different tissues and conditions; we estimate average HERV activities, over the tissues and conditions in their proportions in the database. Tissue and condition-specific activities will be studied later.

[6] To make the generated ESTs “realistic”, the parameters of the generating HMM were close to the parameters learned from real data and the lengths of the ESTs were controlled with a heuristic.

[7] We tested the effect of the two heuristics on a tiny test data set; the heuristics gave (to a very high precision) the same results as the complete model (results not shown).

family. The behavior in other families is very similar. The standard deviations are small compared to the differences between the active and almost inactive HERVs. Thus we can trust the active-looking ones to truly be active.

We expected that active HERVs would be young or otherwise have a well-preserved sequence. However, for this small real data neither intactness nor age of the element seems to visibly explain high activity. The second-most active HERV is among the oldest sequences in the data. Some active HERVs are missing portions of the typical retrovirus sequence, and some full length elements are silent.

Quick search of the public sequence databases shows that not much is known about the top 15 active HERVs. One of the active HERV-W sequences is syncytin [2], a retroviral human gene known to be expressed in the placenta. The second most active HERV is a sequence containing a putative human gene, but it is in an area of the HERV that does not have a known retroviral function according to RetroTector. This merits further investigation. Since activity of a large portion of the HERV sequences was previously unknown, finding their activation is truly new information.

HERV activity is commonly studied for families instead of individual HERVs; this is done by measuring the activity of hand-picked model sequences for the families. It is striking that when we looked at the activities of these model sequences (included in our set of 91 HERVs), they were among the top 15 active HERVs (9th, 12th, 15th), but were not the most active ones within the family.

We compare our method to a simple BLAST based approach with generated data. In the BLAST approach each EST is counted in favor of its best matching HERV and the activities of HERVs are given as counts of ESTs in their favor. The BLAST approach was used also in [4] for a tiny data set containing only intact HERV sequences. Both methods performed about equally; see Fig. 2 (right). The Kullback-Leibler divergence between the learned activity distribution and the generating distribution (ground truth) was 0.035 for our method and 0.036 for the BLAST approach. Both the HMM model and the BLAST approach preserved 18 of the top 20 most active HERVs. The simple approach is surprisingly good compared to HMM-based modeling; this suggests it can be used for large tests where HMM training would be computationally too costly.

4 Conclusions

We have introduced a generative model-based method that estimates the activities of individual HERVs. Such analysis is vital for understanding the underlying control mechanisms of HERV activation. HERVs reported as active with our method can later be verified with laboratory methods; by contrast, exhaustive search of active HERVs with laboratory methods would be too expensive.

In artificially generated data both our method and a heuristic BLAST-based alternative were able to estimate underlying activities fairly well: the top active HERVs were among the top active in the results. This justifies the use of the computationally simpler alternative instead of the rigorous probabilistic method.

In real data, in each HERV family we detected several new active HERVs that need further biological analysis. The generally used model sequences of the families were not the most active, which suggests that the more active HERVs could be better probe sequences in expression studies. No clear relation between age/intactness and activity was visible.

In summary, we have introduced a method for extracting expression levels from sequence databases. The method is general; we can use it to compare HERV activities in different conditions, or to study endogenous retroviruses in other organisms or to study other kinds of transposable elements.

References

1. Nelson, P.N., Carnegie, P.R., Martin, J., Davari Ejtehadi, H., Hooley, P., Roden, D., Rowland-Jones, S., Warren, P., Astley, J., Murray, P.G.: Demystified ... human endogenous retroviruses. Molecular Pathology **56** (2003) 11–18
2. Muir, A., Lever, A., Moffett, A.: Expression and functions of human endogenous retroviruses in the placenta: An update. Placenta **25**(Suppl. 1) (2004) S16–S25
3. Seifarth, W., Frank, O., Zeifelder, U., Spiess, B., Greenwood, A.D., Hehlmann, R., Leib-Mösch, C.: Comprehensive analysis of human endogenous retrovirus transcriptional activity in human tissues with a retrovirus-specific microarray. Journal of Virology **79**(1) (2005) 341–52
4. Stauffer, Y., Theiler, G., Sperisen, P., Lebedev, Y., Jongeneel, C.V.: Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. Cancer Immunity **4**(2) (2004)
5. Kim, T.H., Jeon, Y.J., Kim, W.Y., Kim, H.S.: HESAS: Hervs expression and structure analysis system. Bioinformatics **21**(8) (2005) 1699–1700
6. RetroTector. http://www.kvir.uu.se/RetroTector/RetroTectorProject.html.
7. Expressed Sequence Tags database. http://www.ncbi.nlm.nih.gov/dbEST/.
8. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215** (1990) 403–10